Modeling Recovery Curves for Prostate Cancer

by

Fulton Wang

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Masters of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

Author .....
Department of Electrical Engineering and Computer Science
May 22, 2013

Certified by .....
Cynthia Rudin
Assistant Professor
Thesis Supervisor

Accepted by .....
Leslie Kolodziejski
Chairman, Department Committee on Graduate Theses

Modeling Recovery Curves for Prostate Cancer

by

Fulton Wang

Submitted to the Department of Electrical Engineering and Computer Science
on May 22, 2013, in partial fulfillment of the
requirements for the degree of
Masters of Science

Abstract

In this thesis, I analyze a dataset containing a time series of various bodily functionality scores for patients following 3 different forms of prostate cancer treatment. I propose a parametric model for those function level time series and show promise that the method can be applied to both simulated and real data.

Thesis Supervisor: Cynthia Rudin
Title: Assistant Professor

Acknowledgments

I would like to thank my thesis advisor Cynthia Rudin for her guidance and care, and her willingness to let me join her group. I would like to thank Tyler McCormick for his cheerful collaboration and insights, Jim Michaelson over at MGH for giving us the medical background we sorely lack. The small legion of students at MGH provided much needed levity when the data got difficult to work with. Old labmates - I'm no longer there, but you guys played a big part in my academic development. Good friends - thanks for supporting me thoughout my journey, and finally, I thank my family for their unwavering faith in me.

Contents

1 Introduction **11**

1.1 Background 11

1.2 Goal of Project 12

1.3 Desirables for Model 12

2 Exploratory Data Analysis **13**

2.1 Dataset 13

2.2 General Shape of Function Curves 13

2.3 Dependence of Function Curves on Patient Attributes 14

3 Model Description **17**

3.1 Model Description 17

3.2 Parametric Form of Curve 17

3.3 Model for Curve Parameters 19

3.3.1 Generalized Linear Model 19

3.3.2 Model for a 20

3.3.3 Beta Distribution 20

3.3.4 GLM for a 21

3.3.5 GLM for c 21

3.3.6 Data Normalization and choosing intercepts z^a, z^b, z^c 22

3.3.7 Extracting mean population parameters $\mu^a_{pop}, \mu^b_{pop}, \mu^c_{pop}$ 22

3.4 Model for observed data 23

3.5 Priors for parameters 23

3.5.1 Desired prior predictive distribution for a, b, c 24
3.5.2 Specifying the Priors for Parameters 24
3.5.3 Prior Predictive Distribution for $\tilde{\mu}^a$ 25
3.5.4 Logit-Normal Distribution 25
3.5.5 Logit-Normality of $\tilde{\mu}^a$ in the Prior 27
3.5.6 Dependence of Prior Predictive Distribution of $\tilde{\mu}^a$ on Hyperparameters 27
3.5.7 Choosing c^a and λ^a 29
3.5.8 Prior Predictive Distribution for $\tilde{\mu}^c$ 29
3.5.9 Choosing c^c and λ^c 31
3.6 Plots of prior patient curve distributions 32

4 Curve Prediction with Model **33**
4.1 Bayesian Inference 33
4.2 Simulation Results 33
4.2.1 Simulating latent variables a, b, c 34
4.2.2 Simulating data points $g^*(t)$ 34
4.3 Biasedness of model 34
4.3.1 Applicability of Model to Real Data 37

List of Figures

2-1 Average patient time series for the 3 side effects, stratified by treatment 14

2-2 Average patient time series for the 3 side effects, stratified by each of 6 patient attributes 15

2-3 Average patient time series for the 3 side effects, stratified by each of 6 patient attributes 15

2-4 Initial drop in function level vs attribute, for 3 side effect/treatment combinations 16

3-1 prior predictive distribution of $\tilde{\mu}^a$ 28

3-2 prior predictive distribution of $\tilde{\mu}^c$ for several values of μ^c_{pop} and $\tilde{\sigma}^c$ . . 31

3-3 prior predictive distribution of over curves for several values of $\tilde{X}$ . . 32

4-1 Plots for when simulating a, b, c 35

4-2 Plots for when simulating a, b, c 36

4-3 histogram of posterior of B_a when ϕ^a is fixed to various values during inference 38

4-4 Plots of bowel function curve parameters vs initial function level and age attribute, stratified by treatment 40

4-5 Plots of sexual function curve parameters vs initial function level and age attribute, stratified by treatment 41

4-6 Plots of sexual function curve parameters vs initial function level and age attribute, stratified by treatment 42

Chapter 1

Introduction

1.1 Background

Prostate cancer is one of the most common cancers, affecting roughly 1 in 5 men in the USA. Men diagnosed with PC typically have several treatment choices, and to decide between them, one would need to consider what effect the treatment will have on various bodily functions. For example, radial prostatecomy is known to have a fairly strong effect on sexual function level. Other bodily functions of interest include urinary, bowel, and general physical function. To help the patient make their decision, it would be very helpful to provide them with a prediction of what their various function levels would be, should they choose a given treatment.

Typically, each of these functions can be quantified, and are known to change with time. For example, one can be assigned a sexual function score, and typically one's sexual function score undergoes a steep drop immediately after surgery, before slowly recovering to some steady state level. Then, the piece of information that would be of use to a patient is to offer them, for each treatment and each function, a *time series* of the function level, should they undergo the given treatment.

Furthermore, data shows that one's function time series for a given treatment varies by patient; different patients undergoing the same treatment should expect to, on average, experience *different* function time series. For example, patients who are doing poorly in terms of sexual function before surgery should expect to do worse in

the long run than patients who were sexually unhealthy before surgery.

1.2 Goal of Project

Therefore, our goal is to build a predictive model of the *personalized* function time series for patients, taking into account patient attributes such as age, race, comorbidity, and function level prior to treatment.

To help people make decisions, it is not enough to present to a patient a single most likely estimate of their function time series given a treatment - if one is not very certain in the prediction, then that information should be communicated to the patient. Thus, we will take a Bayesian approach to curve prediction that furthermore utilizes a novel prior structure, reflecting our a priori belief that the more 'extreme' a patient is, the the more uncertainty there would lie in our predictions for a patient.

1.3 Desirables for Model

We would like our model to be Bayesian, because we want to know how much uncertainty there is in our curve predictions. We want the model to be easily intrepretable -the parameters of our model should have easy to understand meanings. Finally, we want to build into our model the belief that the 'average' patient should be predicted to have a curve that is the 'average' of all the curves in the dataset. Without patient-specific predictions, a patient would simply look up a study of prostate cancer, and see on average, how one side effect is affected if he should choose a particular treatment. In other words, a patient would regard himself as being average, and should expect to have the average response to treatment. In the case of patient-specific prediction, we believe the average patient should still map to the average curve. Furthermore, in the absence of data, we should do as before, and predict all patients to have the average response. This belief will be encoded in the prior distribution for our Bayesian model.

Chapter 2

Exploratory Data Analysis

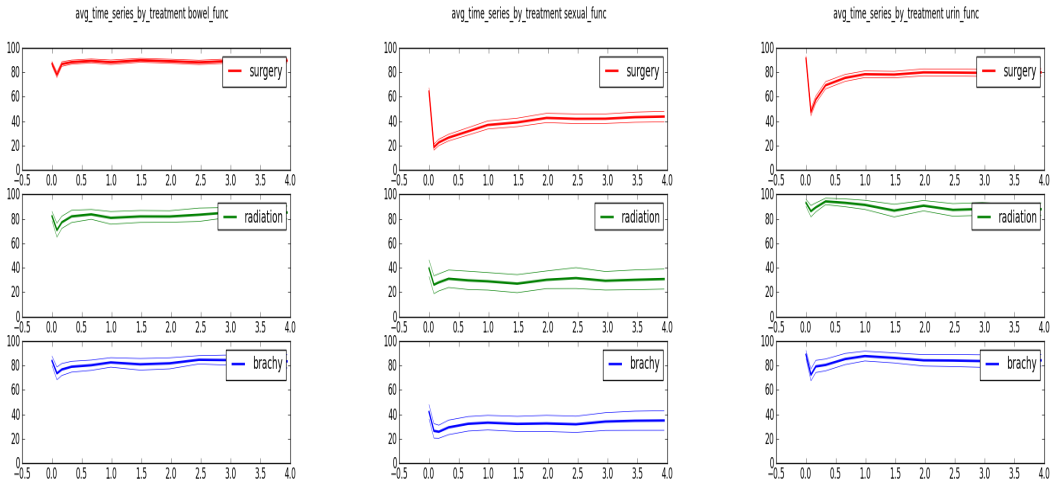
2.1 Dataset

To build our predictive models, we use a dataset collected by urologists at UCLA, which has previously been used in publication[2]. This data follows a cohort of roughly 1000 patients who received one of 3 possible treatments - prostatectomy, radiation therapy, or brachytherapy for a period of 5 years. Surveys were sent at various time points after treatment that asked patients to assign a functional score in each of several categories: sexual, bowel, and urinary function, as well as general physical and mental well being. These scores are between 0 and 100. Several patient attributes such as age, race, comorbidity count, and PSA level were also recorded for every patient. Figure 1 shows an example of the average sexual function scores in the entire dataset after each of 3 treatments.

2.2 General Shape of Function Curves

The very first thing we did was to see on average what the function curves looked like for different patients, and whether they differed by treatment. Below, for each of the 3 side effects, we plot the aggregate function time series for patients opting for each treatment.

All of the aggregate curves seem to have a similar shape: an initial instantaneous



(a) bowel function (b) sexual function (c) urinary function

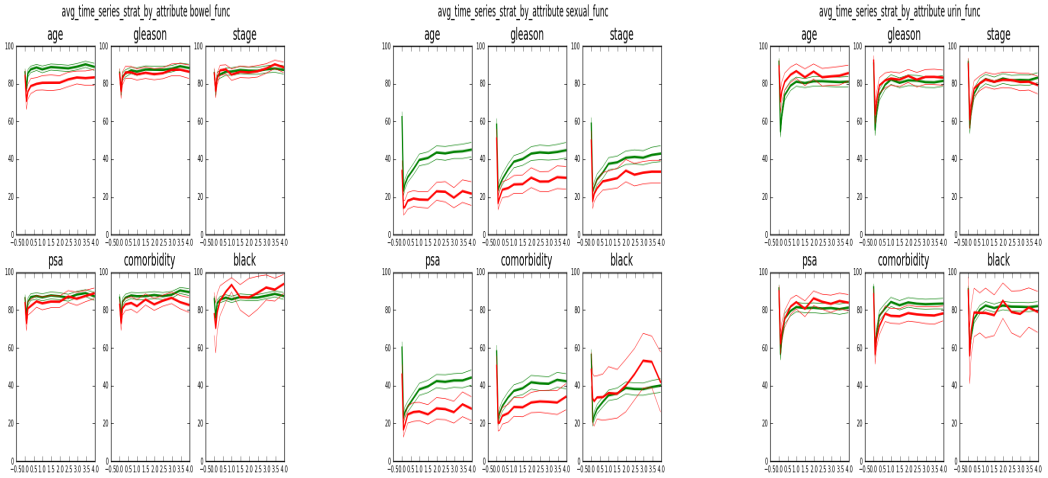
Figure 2-1: Average patient time series for the 3 side effects, stratified by treatment

drop off in function level, followed by a rise to some steady state function level. This hints that we should model the curves parametrically, parameterizing the key attributes of it - the initial drop, the long-term drop, and the rate of function recovery. Secondly, the treatment chosen does seem to affect the level of initial function drop and long term function level.

2.3 Dependence of Function Curves on Patient Attributes

The second thing we wanted to look at was for a given side effect and treatment, whether the function curves vary depending on the available attributes. Below, for each of the 3 side effects, for each of the 6 attributes we have for patients, divide the dataset into 2 halves, based on the given attribute. We plot the average function time series for each half of the dataset, to see if there is a difference.

It seems like there is a difference in the curves, depending on various attributes. However, the attributes seem to be strongly correlated with the pre-treatment function level. It seems logical that the pre-treatment state would be highly correlated with the post-treatment function level. To verify this, we make the same plots as be-

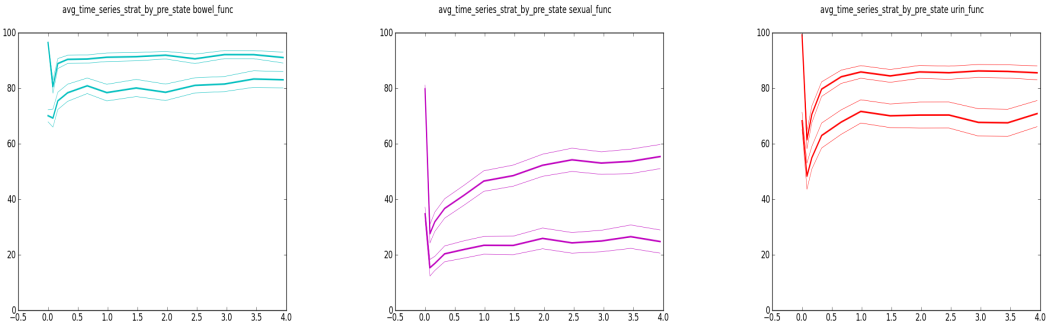


(a) bowel function (b) sexual function (c) urinary function

Figure 2-2: Average patient time series for the 3 side effects, stratified by each of 6 patient attributes

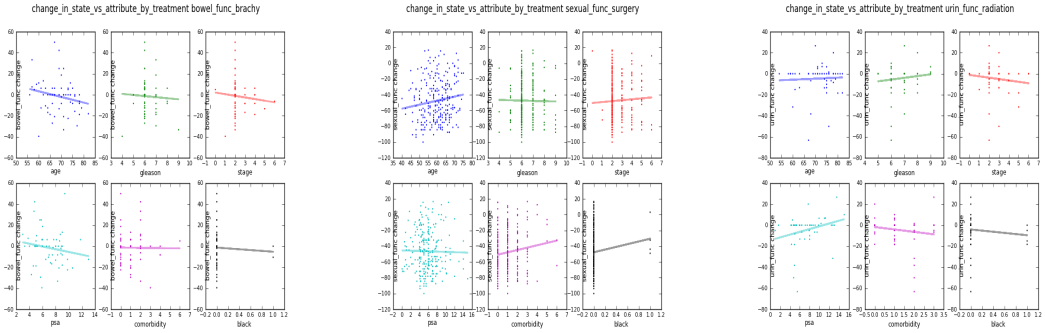
fore, except this time, we stratify the patients by their pre-treatment function level. Indeed, pre and post-treatment function levels are highly correlated.

We wonder whether after controlling for the pre-treatment state, a patient's attributes still impact their curves. Thus, for each side effect, treatment combination, for each of the 6 attributes, we made a scatter plot of the attribute vs the change in side effect function level before treatment and right after treatment (at the 1 month survey time). There are too many side effect/treatment combinations to display scatter plots for, but below, we choose 3 such combinations and show their associated



(a) bowel function (b) sexual function (c) urinary function

Figure 2-3: Average patient time series for the 3 side effects, stratified by each of 6 patient attributes



(a) bowel function/brachytherapy (b) sexual function/surgery (c) urinary function/radiation

Figure 2-4: Initial drop in function level vs attribute, for 3 side effect/treatment combinations

scatter plots.

There is definitely no trend for some attributes, but perhaps for some there is a slight one.

Chapter 3

Model Description

3.1 Model Description

Here, we describe the model we will use to model side effect function values. First, we note that we will use completely independent models for each side effect and treatment combination. While we did not have to adopt such an approach (for example, a patient's function might have a patient-specific and treatment-specific component), we do so for simplicity's sake.

In short, we assume that each patient has a 'true' function curve $g_i(t)$, and that the observed function values $g_i^*(t_j)$ are normally distributed about $g_i(t_j)$. The 'true' function curve will be parameterized by 3 latent parameters, each of which is modelled with a generalized linear model that depends on patient covariates. In the prior, the parameters for these 3 geenralized linear models are independent, though in the posterior, they will become dependent through the observed function values.

3.2 Parametric Form of Curve

Recalling the general function curve shapes from the previous section, we decide the 3 things we want to model of the curve $g_i(t)$ are:

1. The long term drop in function value

2. The short term drop in function value

3. The rate at which the function value recovers from the initial drop to the long term value.

Furthermore, we place the following restrictions on the curve:

1. The long term drop in function value is indeed a drop; that is, the long term function value is less than the pre-treatment function value

2. The short term drop is greater than the long term drop in function value.

3. 'True' function values at all times are between 0 and 1.

With these considerations in mind, we are ready to describe the parameterization of $g(t)$:

$$
\begin{aligned}
g(t; s, a, b, c) &= s(1 - a - b(1 - a)e^{-ct}) && (3.1)\\
\text{where} & && (3.2)\\
a &\in (0, 1) && (3.3)\\
b &\in (0, 1) && (3.4)\\
c &\in (0, \infty) && (3.5)\\
& && (3.6)
\end{aligned}
$$

and

- s is the pre-treatment function level

- a is the long term loss in function level, relative to the pre-treatment function level

- b is the short term loss in function level in excess of the long term loss, expressed as a portion of the long term function level

- c is the rate of the function level decays from the short term to long term function level

Note that the restrictions we place on the curve are enforced by the restrictions on a, b, c.

3.3 Model for Curve Parameters

Now that we defined how the 'true' function level curve is parameterized, we will define the models for each of the parameters. As the parameters' ranges are constrained, we cannot use a standard linear model for the parameters. However, generalized linear models will suit us perfectly. Furthermore, as a and b have the same range, the models for a and b will be analogous to each other.

3.3.1 Generalized Linear Model

Standard linear regression models the observed variable Y as coming from a Normal distribution such that $E(Y)$ is a linear function of the covariate vector X. That is, $Y \sim N(BX, \sigma)$, where σ is the standard deviation of Y, and more generally, a parameter that either directly or indirectly specifies the spread of Y. Generalized linear regression generalizes linear regression in 2 ways: 1. The observed variable Y comes from a distribution with a density $p(Y)$ that is not necessarily that of the normal distribution, and 2. $E(Y)$ is no longer necessarily a linear function of covariates, but the result of a linear function of covariates that is subsequently sent through a link function. That is, $E(Y) = f(BX)$, where f is the link function. Specification of f allows one to control the manner in which the mean response $E(Y)$ depends on covariates. In particular, it allows one to restrict the range of $E(Y)$.

A Generalized linear model relating the observed variable Y to covariate vector X thus has 2 components:

- a probability distribution $p(Y)$ parameterized, its mean $\mu = E(Y)$ and a parameter ϕ controlling the variance of Y.

- a link function f such that $\mu = f(BX + z)$

Note that we explicitly allow for a bias term z. For example, linear regression can be thought of as a generalized linear model where $p(Y) = p_{\text{normal}}(y; \mu = \mu, \sigma = \phi)$, and the link function f being the identity function.

3.3.2 Model for a

As a must reside in the unit interval, if we are to model a using a GLM, $p(a : \mu, \phi)$ must have zero support outside the unit interval. The Beta distribution is a good candidate for $p(a)$. In fact, GLM using a beta expression to model the dependent variable has been studied in detail before[1].

3.3.3 Beta Distribution

This is a continuous distribution with support in (0,1). There are various parameterizations of the Beta distribution. The most common one, for a distribution $Beta(\alpha, \beta)$ is as such:

$$p_{Beta}(y; \alpha, \beta) \propto y^{\alpha-1}(1-y)^{\beta-1} \tag{3.7}$$

with $E(X) = \frac{\alpha}{\alpha+\beta}$ and $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. However, for GLM, we require $p(Y)$ to be parameterized by $\mu = E(Y)$ and some dispersion parameter ϕ. Fortunately, some properties of the Beta distribution allow for such a parameterization. Firstly, if we let

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{3.8}$$

$$\phi = \frac{1}{\alpha + \beta + 1} \tag{3.9}$$

we see that a $Beta(\alpha, \beta)$ random variable has mean μ and variance $\mu(1-\mu)\phi$. We can then solve the above 2 equations for μ and ϕ in terms of α and β. Then, we can

obtain an alternate parameterization for a Beta random variable $p(y; \mu, \phi) =$

3.3.4 GLM for a

Now, we can describe the model for a_i concisely:

$$a_i \sim Beta(\mu_i^a, \phi^a) \tag{3.10}$$

$$\mu_i^a = f^a(B^a X_i + z^a) \tag{3.11}$$

$$f^a(x) = \frac{1}{1 + e^{-x}} \tag{3.12}$$

f^a is the link function for the GLM, which we have chosen to be the logistic function, as μ_i^a must be between 0 and 1. We have not described what z^a will be yet, but will do so shortly. b is defined analogously to a. We have not defined what distribution ϕ^a follows. For now, the important thing is that ϕ^a is shared between patients, and is between 0 and 1.

3.3.5 GLM for c

Keeping in mind that $c > 0$, we will let c come from a gamma distribution. The gamma distribution can be parameterized by its mean μ and a shape parameter k that controls the variance of the distribution. We would like c to come from a unimodal distribution, which will be the case if $k > 1$. We model the inverse of k instead, letting $\phi^c = \frac{1}{k}$ in the traditional gamma distribution parameterization, keeping in mind that k is between 0 and 1, and is again, shared between patients.

Thus, the model for c_i is as such:

$$c_i \sim Gamma(\mu_i^c, \phi^c) \tag{3.13}$$

$$\mu_i^c \sim f^c(B^c X_i + z^c) \tag{3.14}$$

$$f^c(x) = e^x \tag{3.15}$$

Once again, we defer specification of the priors for the parameters z^c, B^c and ϕ^c

until later.

3.3.6 Data Normalization and choosing intercepts z^a, z^b, z^c

The way we normalize the data and choose the intercept parameters is guided by our desire in the model that the the 'average' patient should, under our model, expect to receive the average curve as their 'true' curve, which is a curve for which a, b, c are equal to the average value of those parameters in our dataset, which we will denote as $\mu^a_{pop}, \mu^b_{pop}, \mu^c_{pop}$, respectively. We would like, for the patient with the average covariate vector X_{pop}, $E(a) = \mu^a_{pop}$, and likewise for b and c. We can accomplish by doing the following:

1. normalize each covariate to have 0 mean and standard deviation 1 across the dataset
2. Set $z^a = g^{a-1}(\mu^a_{pop})$

Looking at the equation for μ^a_i, one sees that this way, the average patient has a covariate vector equal to the 0 vector, and receives a value of $\mu^a_i = \mu^a_{pop}$.

3.3.7 Extracting mean population parameters $\mu^a_{pop}, \mu^b_{pop}, \mu^c_{pop}$

To calculate these mean parameters, we need to know their values for each patient. But we don't actually observe patient curve parameters a, b, c, and so we will obtain them through least squares curve fitting, ensuring that the curve parameters obey their constraints. More specifically, for each patient i, we want to:

$$\begin{aligned} \text{minimize} \sum_{j=1}^{m} (g^*_i(t_j) &- g_i(t_j; s_i, a_i, b_i, c_i))^2 \\ \text{subject to} & \\ a_i &\in (0, 1) \\ b_i &\in (0, 1) \\ c_i &\in (0, \infty) \end{aligned}$$

For curves where the function value does not drop (or even increases) after treatment, a will be quite small or even 0, and the fitted curve may be flat when the actual curve rises. However, we view this as unavoidable due to the restrictions of the kinds of curve we allow in our model.

3.4 Model for observed data

We have described how a patient's true curve $g(t; s, a, b, c$depends on 3 latent parameters a, b, c and observed pre-treatment value s, and how those 3 parameters depend on patient covariates and the parameters of the model B_a, B_b, B_c. Now, it remains to specify how the observed data $g^*(t_j; s, a, b, c)$ depends on $g(t_j; a, b, c)$. We will take a simplistic model. We will assume that the observed function value at time t_j is normally distributed about the true function value. That is,

$$g_i^*(t_j; s_i, a_i, b_i, c_i) \sim N(g_i(t_j; s_i, a_i, b_i, c_i, \sigma^{noise}) \tag{3.16}$$

for all patients i and at all measurement times $t_1, \ldots, t_j$.

Thus, we are assuming that the observed function values for a patient are conditionally independent of each other given the patient's true curve parameters. In reality perhaps those observed values might be correlated, but we opt for simplicity whenever possible.

3.5 Priors for parameters

Thus far, our model contains 7 parameters: $\theta = \{B_a, B_b, B_c, \phi^a, \phi^b, \phi^c, \phi^{noise}\}$. As we adopt a Bayesian framework, we must give prior distributions for each of those parameters $P(\theta; \alpha)$ where α is a set of hyperparamters.

3.5.1 Desired prior predictive distribution for a, b, c

To describe what properties we want of $P(\theta; \alpha)$, it is more useful to describe, before observing any data, what we want the distribution over $g(t; s, a, b, c)$ to look like. As $g(t; s, a, b, c)$ is fully described by a, b, c, then what we want to do is, for an test patient X, describe the prior predictive distributions $P(a; \alpha, X), P(b; \alpha, X), P(c; \alpha, X)$. X denotes the test sample, not the data that we have yet to observe. These distributions will of course depend on X. We want $P(a; \alpha, X)$ to be:

1. unimodal, for reasonably values of X
2. roughly centered around μ^a_{pop}
3. to have larger variance the further X is from the average covariate vector. (since we are normalizing the data to be mean 0, this means the larger the magnitude of X, the larger the variance of $P(a; \alpha, X)$.

These same desirables apply to the prior predictive distributions of b and c.

3.5.2 Specifying the Priors for Parameters

We will let the parameters B_a, B_b, B_c each follow normal distributions in the prior. These normal distributions will be be mean zero, and have a diagonal covariance matrix equal to some scalar multiple of the identity matrix. That is, we will let:

$$B_a \sim N(0, c^a I) \tag{3.17}$$
$$B_b \sim N(0, c^b I) \tag{3.18}$$
$$B_c \sim N(0, c^c I) \tag{3.19}$$
$$\tag{3.20}$$

where c^a, c^b, c^c are hyperparameters, as they describe prior distributions of model parameters. We will also let the disperson parameters $\phi^a, \phi^b, \phi^c, \phi^{noise}$ follow exponential distributions truncated at 1. That is, we will let:

$$\phi^a \sim \text{truncated\_exp}(\lambda^a, 1) \tag{3.21}$$

$$\phi^b \sim \text{truncated\_exp}(\lambda^b, 1) \tag{3.22}$$

$$\phi^c \sim \text{truncated\_exp}(\lambda^c, 1) \tag{3.23}$$

$$\phi^{noise} \sim \text{truncated\_exp}(\lambda^{noise}, 1) \tag{3.24}$$

$$\tag{3.25}$$

The reason why we choose the prior for these parameters is that we want to encourage them to be small so that the distributions for a, b, c will have relatively small variance in the prior. The reason for this is that in general, distributions with high variances with finite support will not be unimodal, which is a situation we want to avoid.

Now, we analyze the influence of the hyperparameters c^a and λ^a on the prior predictive distribution $P(X|\alpha)$. We will do so in 2 steps:

1. Study how $P(\mu^a; c^a, X)$ depends on X and c^a and choose c^a.

2. Study how $P(a; c^a, X, \lambda^a)$ depends on λ^a, once c^a is chosen

3.5.3 Prior Predictive Distribution for $\tilde{\mu}^a$

Here, we study the prior predictive distribution of $\tilde{\mu}^a$, the underlying 'true' value of the parameter a for a patient with covariate vector $\tilde{x}$. The goal of this section is to see how our prior belief on what $\tilde{\mu}^a$ varies depends on the hyperparameters. We first claim that in the prior predictive distribution, $\tilde{\mu}^a$ follows a logit-normal distribution, whose properties we will now describe.

3.5.4 Logit-Normal Distribution

A logit-normal distribution is a continuous distribution defined on the open interval $(0, 1)$. A random variable X follows a logit-normal distribution if the transformed

random variable logit(X) follows a normal distribution. An equivalent definition makes the parameterization of a logit-normal distribution clear:

Definition 1. If $Y \sim \text{Normal}(\mu, \sigma)$, then the transformed random variable $X = \text{logistic}(Y)$ follows a Logit-Normal(μ, σ) distribution.

Thus, a Logit-Normal distributed random variable X is parameterized using the mean and standard deviation of the normal distribution that Logit(X) follows.

Key Properties

Applying the change of variable formula to the density function of a Normal(μ, σ) random variable under the logistic transformation, one arrives at the density function of a Logit-Normal(μ, σ) random variable:

$$f_X(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(\text{logit}(x) - \mu)^2}{2\sigma^2} \frac{1}{x(1-x)}; x \in (0, 1) \tag{3.26}$$

Unfortunately, there are no analytical formulas for the mean, variance, or mode of a logit-normal distribution.

Unimodality

In general, the density of a logit-normal distribution may have either 1 or 2 modes. For sufficiently large σ, a Logit-Normal(μ, σ) distribution will have 2 modes - one near 0 and one near 1. This intuitively makes sense, because a sufficiently diffuse normal distribution will have significant mass far away from 0, where the slope of the logistic function is nearly flat. Conversely, for sufficiently small σ, $f_X(x; \mu, \sigma)$ will be unimodal. This result can be seen analytically.

By taking the derivative of $f_X(x; \mu, \sigma)$, it can be seen that the modes of $f_X(x; \mu, \sigma)$ occur at x for which the following condition is satisfied:

$$\text{logit}(x) = \sigma^2(2x - 1) + \mu \tag{3.27}$$

For any μ and σ, there is always at least 1 x for which this condition is satisfied. Thus, $f_X(x;\mu,\sigma)$ always has at least 1 mode. Furthermore, as the slope of logit(x) is always greater than 1, we see that if $\sigma^2 < 1$, there is exactly 1 x such that the condition is satisfied, and arrive at the following observation:

Observation 1. If $\sigma^2 \leq 1$, then $f_X(x;\mu,\sigma)$ is unimodal.

3.5.5 Logit-Normality of $\tilde{\mu}^a$ in the Prior

The only hyperparameter that $\tilde{\mu}^a$ depends on in the prior is $\Sigma^a = c^a I$. Now, we see that $\tilde{\mu}^a|c^a$, the prior predictive distribution of μ^a, follows the logit-normal distribution. This is because $B^a \sim N(0,\Sigma^a)$ and so $B^a\tilde{x} \sim N(0,\tilde{x}'\Sigma^a\tilde{x})$. Then, $\mu^{a*}_{pop} + B^a\tilde{x} \sim N(\mu^{a*}_{pop},\tilde{x}'\Sigma^a\tilde{x})$. Finally, as $\tilde{\mu^a}|c^a \sim g^a(\mu^{a*}_{pop} + B^a\tilde{x})$ and g^a was defined to be the logistic function, we arrive at the following observation:

Observation 2. $\tilde{\mu}^a|c^a \sim$ Logit-Normal$(\mu^{a*}_{pop},\tilde{\sigma}^a)$, where $\tilde{\sigma}^a = \tilde{x}'\Sigma^a\tilde{x} = c^a\sum_{j=1}^k \tilde{x}_j^2$

where we have used the fact that Σ^a was parameterized by the hyperparameter c^a.

3.5.6 Dependence of Prior Predictive Distribution of $\tilde{\mu}^a$ on Hyperparameters

For a test sample $\tilde{x}$, the prior predictive distribution $\tilde{\mu}^a|c^a$ is distributed Logit-Normal$(\mu^{a*}_{pop},\tilde{\sigma}^a)$, where $\mu^{a*}_{pop} = g^a(\mu^a_{pop})$ and $\tilde{\sigma}^a$ is as defined in the previous section. Let us first see how this prior predictive distribution depends on μ^a_{pop} and $\tilde{\sigma}^a$. In Figure X, we have plotted for several values of μ^a_{pop}, how the prior predictive distribution $\tilde{\mu}^a|c^a$ depends on $\tilde{\sigma}^a$. The vertical line denotes the location of μ^a_{pop}.

There are several things to note from the figure:

1. As $\tilde{\sigma}^a$ approaches 0, Logit-Normal$(\mu^{a*}_{pop},\tilde{\sigma}^a)$ converges to the point mass at μ^a_{pop}. Recall that we are normalizing the covariate vectors so that each covariate has mean 0 and standard deviation 1 over the training data, so that if a test sample

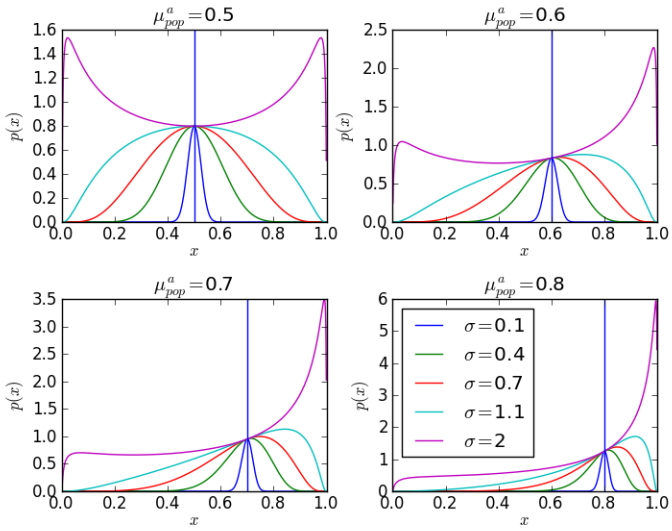


Figure 3-1: prior predictive distribution of $\tilde{\mu}^a$

$\tilde{x}$ is equal to the 'average' patient of the training data, then $\tilde{x} = 0$. This means that the more similar a test sample $\tilde{x}$ is to the 'average' patient, the more closer $\tilde{x}$ is to 0, the smaller $\tilde{\sigma}^a$ is, and thus the more strongly we believe in the prior that $\tilde{\mu}^a$ is equal to μ^a_{pop}, the average of a in the training data. This is what we want.

2. If $\tilde{\sigma}^a \leq 1$, then $P(\tilde{\mu}^a; c^a)$ is necessarily unimodal. If $\tilde{\sigma}^a > 1$, then $f(\tilde{\mu}^a; c^a)$ might still be unimodal, but not necessarily.

3. As $\tilde{\sigma}^a$ increases, the mode of $P(\tilde{\mu}^a; c^a)$ increases. While we would like the mode to remain constant as $\tilde{\sigma}^a$ increases, we see this as being unavoidable. Fortunately, the spread of $P(\tilde{\mu}^a; c^a)$ also increases, so that we have a weaker prior belief over $\tilde{\sigma}^a$.

4. The mean of $\tilde{\mu}^a | c^a$ decreases as $\tilde{\sigma}^a$ increases. So the mode and mean exhibit opposite trends.

In Figure X, we plot how the mode of $P(\tilde{\mu}^a; c^a)$ changes with $\tilde{\sigma}^a$, as μ^a_{pop} is held fixed. We do this for several values of μ^a_{pop}.

3.5.7 Choosing c^a and λ^a

We are ultimately concerned in the prior predictive distribution of $\tilde{a}$. The previous analysis informs us that we should choose c^a to ensure that $\tilde{\sigma}^a$ is less than 1 most of the time, as mu^a is unimodal in those situations. Due to our data renormalization, we can make the assumption that $\tilde{X}$ follows a multivariate $N(0, I)$ distribution. Then, to pick c^a, one can follow the following steps:

1. Choose a proportion ρ such that you want proportion ρ of possible test samples $\tilde{X}$ to have an unimodal distribution for $P(\mu^i|C^a, X)$

2. Calculate (analytically or otherwise) the c^a such that $P(\tilde{\sigma}^a < 1) = \rho$.

Once c^a is chosen, we understand the distribution $P(\mu^a; X, c^a)$. However, what we care about is $P(a; X, c^a, \lambda^a)$. Even if $P(\mu^a; c^a, X)$ is unimodal, $P(a; c^a, X, \lambda^a)$ may not be unimodal if λ^a is too small. The reason is that if ϕ^a is too large, the variance of $P(a; c^a, X, \lambda^a)$ will be large even if the variance of $P(\mu^a; X, c^a)$ is small, and large variances preclude unimodality of distributions. While we cannot fix ϕ^a to be small, we can place a prior on ϕ^a that encourages it to be small. As ϕ^a follows a (truncated) exponential distribution with rate parameter λ^a, the larger λ^a is, the more likely ϕ^a is to be small. Thus, we choose λ^a finding a value such that in $P(a; c^a, X, \lambda^a)$ is unimodal most of the time.

3.5.8 Prior Predictive Distribution for $\tilde{\mu}^c$

Here, we study the prior predictive distribution of $\tilde{\mu}^c$, the underlying 'true' value of the parameter c for a patient with covariate vector $\tilde{x}$. The goal of this section is to see how our prior belief on what $\tilde{\mu}^c$ varies depends on the hyperparameters. We first describe the log-normal distribution, because it turns out this is the distribution that $\tilde{\mu}^c$ follows a log-normal distribution in the prior.

Log-Normal Distribution

A log-normal distribution is a continuous distribution defined on the open interval $(0, \infty)$. A random variable X follows a log-normal distribution if the transformed random variable $\log(X)$ follows a normal distribution. An equivalent definition makes the parameterization of a log-normal distribution clear:

Definition 2. If $Y \sim \text{Normal}(\mu, \sigma^2)$, then the transformed random variable $X = \exp(Y)$ follows a Log-Normal(μ, σ^2) distribution.

Thus, a Log-Normal distributed random variable X is parameterized using the mean and standard deviation of the normal distribution that $\log(X)$ follows.

Key Properties

The density of a Log-Normal(μ, σ^2) distribution is:

$$f_X(x; \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp - \frac{(\log(x) - \mu)^2}{2\sigma^2}; x > 0 \tag{3.28}$$

The log-normal distribution is unimodal regardless of the choice of parameters. This is suitable for our poses. There are analytical formulas for the mean and mode; the mean is $\exp(\mu + \frac{\sigma^2}{2})$, and the mode is $\exp(\mu - \sigma^2)$. Like in the case of the logit-normal distribution, neither the mean nor mode are constant as σ^2 increases, for fixed μ. Also, the mode and mean exhibit opposite trends. Finally, as one would expect, the variance is increasing in σ^2.

Log-Normality of $\tilde{\mu}^c$ in the Prior

The only hyperparameter that $\tilde{\mu}^c$ depends on is c^c. The argument for the log-normality of $\tilde{\mu}^c$ is exactly analogous to that for the logit-normality of $\tilde{\mu}^a$. The prior predictive distribution of μ^c, $\mu^c | c^c$, is distributed $g^c(\mu^{c*}_{pop} + B^c\tilde{x})$. $B^c \sim N(0, c^c I)$, and so $B^c\tilde{x} \sim N(0, \tilde{\sigma}^c)$ where $\tilde{\sigma^c} := \tilde{x}'(c^c I)\tilde{x} = c^c \sum_{j=1}^{k} \tilde{x}_j^2$. Thus $\mu^{c*}_{pop} + B^c\tilde{x} \sim N(\mu^{c*}_{pop}, \tilde{\sigma}^c)$ and thus $\tilde{\mu}^c \sim \text{log-normal}(\mu^{c*}_{pop}, \tilde{\sigma}^c)$.

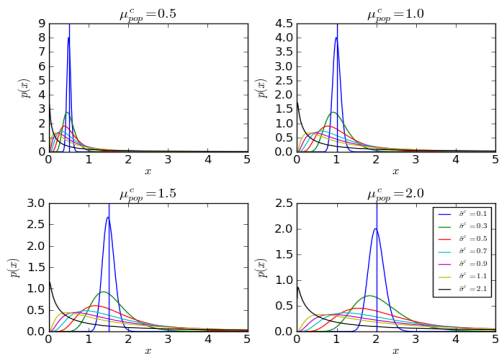


Figure 3-2: prior predictive distribution of $\tilde{\mu}^c$ for several values of μ^c_{pop} and $\tilde{\sigma}^c$

3.5.9 Choosing c^c and λ^c

As $P(\mu^c; c^c)$ is always unimodal, the only consideration we choosing c^c is that if it is too large, the mean of $P(\mu^c; c^c)$ will shift too much as $\tilde{\sigma^c}$ changes. We plotted the distribution of $P(\mu^c; c^c)$ for several values of μ^c_{pop} and $\tilde{\sigma}^c$. From these plots, it seems that setting $c^c = 1$ gives a distribution of $P(\mu^c; c^c)$ that does not shift too much with $\tilde{\sigma}^c$.

To choose λ^c, we once again need to encourage ϕ^c to be small. We found that a value of $\lambda^c = 1$, along with $c^c = 1$, resulted in unimodal distributions for $P(\tilde{c}|c^c, \lambda^c)$.

Choosing a prior for ϕ^{noise}

For the other parameters, we had to choose their corresponding hyperparmeters carefully because we desired that the prior of those parameters be unimodal. However we don't run into such problems with choosing ϕ^{noise}, as we know if the prior on $g(t; s, a, b, c)$ is unimodal, $g^*(g; s, a, b, c)$ will be unimodal regardless of the value of ϕ^{noise}. Then, we are free to estimate the hyperparameter for ϕ^{noise} from the data. We can do so through maximum-likelihood estimation, jointly maximum all variables not conditioned on in the posterior, and seeing what the value of λ^{noise} is.

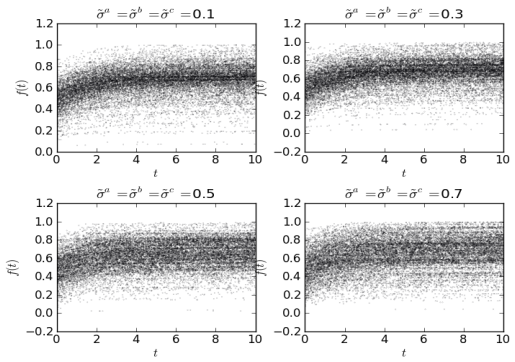


Figure 3-3: prior predictive distribution of over curves for several values of $\tilde{X}$

3.6 Plots of prior patient curve distributions

Having established priors on all parameters by choosing values for the parameters, we have a prior predictive distribution ($P(a, b, c; \alpha, \tilde{X})$ where α denotes the hyperparameters. Thus, we can plot the induced prior predictive distribution over 'true' curves, $P(g(t); s, a, b, c, \tilde{X})$, which we do so below, for several values of $\tilde{X}$. Note that the curves are all centered approximately the same; the mode of the curve distribution is the same regardless of $\tilde{X}$. However, their spread differs.

Chapter 4

Curve Prediction with Model

4.1 Bayesian Inference

The Bayesian approach [3]is simple and concise, and after training, ultimately allows us to generate, for a test sample $\tilde{X}$, a distribution over curves $g(t; s, a, b, c, X)$. In the previous section, we have described a joint distribution over all parameters and data, $P(X, \theta; \alpha)$, where X denotes the observed function value points, and we have integrated out the latent patient curve parameters a, b, c. A patient's curve parameters depend on no other variables besides the parameters θ, if none of the patient's function values are observed, as will be the case when performing prediction. Thus for a test sample, we need to determine $P(\theta; X, \alpha)$. Once we have that, we can directly calculate $P(\tilde{a}; X, \alpha) = P(\tilde{a}; \theta) * P(\theta : X, \alpha)$, and likewise for $\tilde{b}$ and $\tilde{c}$.

To perform the actual inference, we use the standard Metropolis-hastings method[4], with our proposals consisting of cycling through the variables of the distribution and proposing to add a normally distributed noise to it. We use the PYMC package.

4.2 Simulation Results

To show that we can perform posterior inference to extract the posterior distribution of the model parameters B_a, B_b, B_c, we simulated data, fixing B_a, B_b, B_c. We simulated 2 different sets of variables. In both cases, we set $B_a = -1, B_b = 1, B_c = 2$.

Also, we assume the presence of only 1 covariate, and generated 15 covariates equally spaced in the interval (-2,2) to use as the data X.

4.2.1 Simulating latent variables a, b, c

In the first scenario, we set $\phi^a = \phi^b = 0.5$ and $\phi^c = 0.2$, and for each X_i, generated a_i, b_i, c_i from the distribution specified by the model. For example, we generated a_i from a $Beta(B_a * X_i, \phi^a)$ distribution. This model does not contain any actual function values, since a, b, c are directly simulated/observed. To perform inference, we used the same model, fixing ϕ^a, ϕ^b, ϕ^c to the values used to generate a_i, b_i, c_i, and inferred the distribution of $P(B_a|a_i, b_i, c_i, \phi^a, \phi^b, \phi^c, X_i)$, and likewise for B_b and B_c.

4.2.2 Simulating data points $g^*(t)$

In the second scenario, we fix the ϕ^a, ϕ^b, ϕ^c and B_a, B_b, B_c as before. However, we do not simulate a, b, c directly. Rather, we picked a set of times $t_1 \dots t_m$, and for each of the 15 patients, simulated $g_i^*(t_j)$ according to the model. We set $\phi^{noise} = 0.1$. That is, we first simulate a_i, b_i, c_i and once those are determined, simulate $g_i^*(t_j)$ for each time point t_j. To perform inference of B_a, B_b, B_c, we once again assume we know all noise parameters $\phi^a, \phi^b, \phi^c, \phi^{noise}$, and perform sampling to get the distribution of $P(B_a|\{g_i^*(t_j)\}, \phi^a, \phi^b, \phi^c, \phi^{noise}, X_i)$.

4.3 Biasedness of model

From the previous 2 sections, we see that the posterior distributions for B_a, B_b, B_c is actually not centered correctly. To diagnosis this, I created a simplified model to isolate what was going wrong. This is basically the part of the model that predicts a:

$$f^a(x) = \frac{1}{1 + e^{-x}} \tag{4.1}$$

$$\mu_i^a = f^a(\mu_{pop}^a + B^a x_i) \tag{4.2}$$

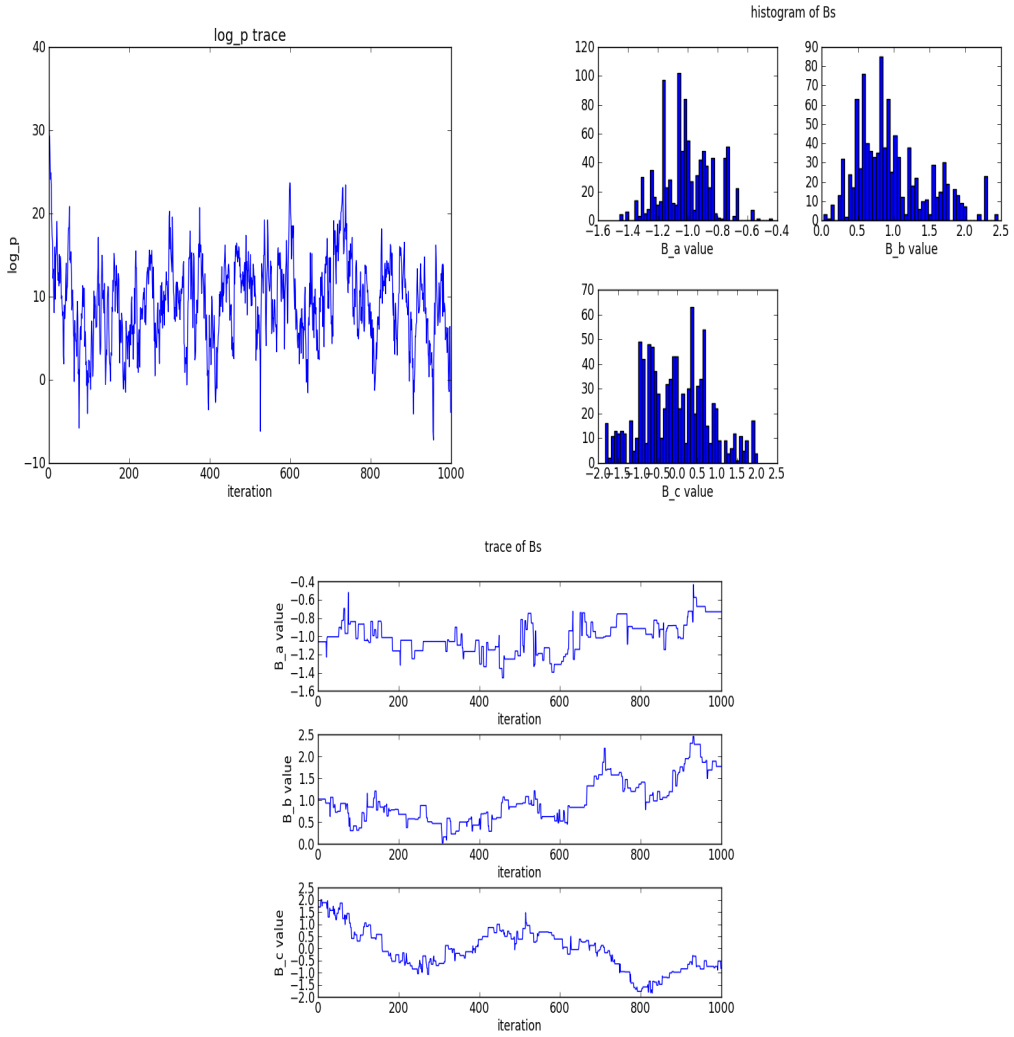


Figure 4-1: Plots for when simulating a, b, c

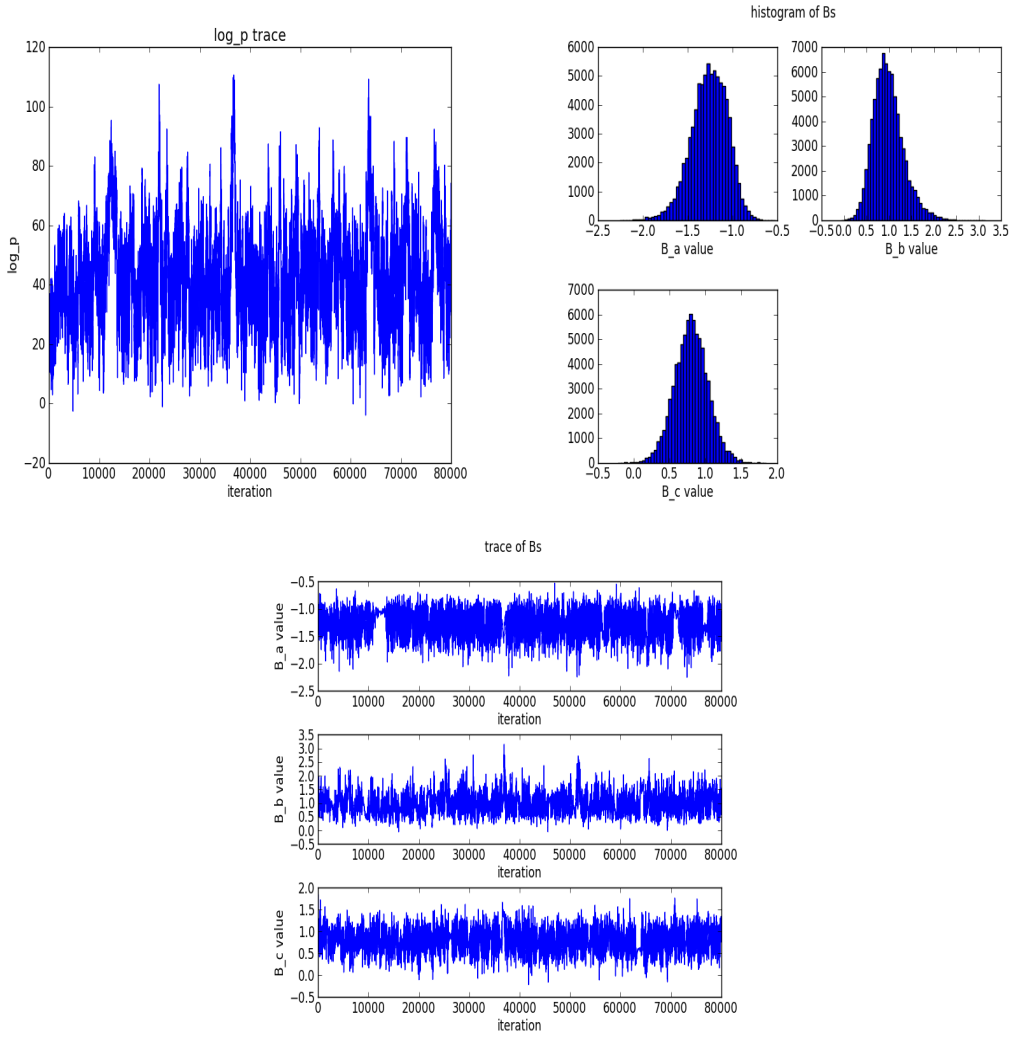


Figure 4-2: Plots for when simulating a, b, c

$$a_i \sim Beta(\mu_i^a, \phi^a) \tag{4.3}$$

$$B \sim U(-\infty, \infty) \tag{4.4}$$

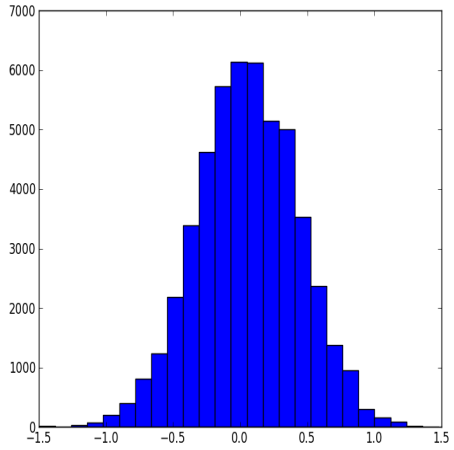
I generate 50 x_i's centered symmetrically about 0. I set $B_a = 1$. I obtain the corresponding a_i's, with no noise, so that $a_i = \mu_i^a$. I assume ϕ^a is fixed. I do sampling to infer $P(B_a|a_i, x_i, \phi^a)$. There is only 1 unobserved variable in this distribution - B_a. The distribution of $P(B_a|a_i, x_i, \phi^a)$ changes as I vary ϕ^a. The distribution is centered correctly at 1.0 for small ϕ^a. The larger ϕ^a is, the more offset the distribution is. See the following plot.

The reason for this biasedness is that if we fix ϕ^a to be large, the distribution for a_i will be U-shaped. I think posterior for B_a is centered at 0 if ϕ^a is close to 1 because when the distribution of a_i is U-shaped, it pays to be off in the predictions.

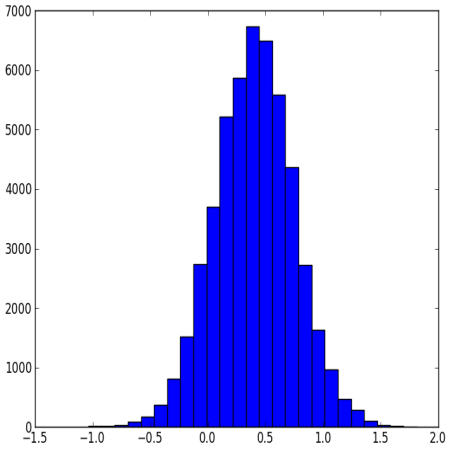
The solution is to give zero probability to situations where the distribution of $P(a; \mu^a, \phi^a)$ is U-shaped. This can be done by parameterizing the Beta distributions for a and b differently. Before, ϕ^a represented the proportion of the maximum possible variance for a Beta distribution with the specified mean. Now, we should let ϕ^a represented the proportion of the maximum possible variance for a Beta distribution with the specified mean, such that the distribution is still unimodal. Hopefully this quantity can be calculated analytically.

4.3.1 Applicability of Model to Real Data

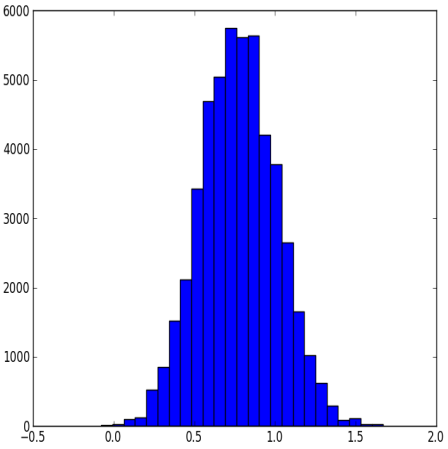
Now that we have parameterized patient function curves, we can explore the relationship between various covariates and curve parameters. Recall from previous plots that the covariates that seem to affect the curve shape the most are age and the pre-treatment function level. On the following few pages, for each of the 3 side effect function values, and for each of treatments, and for each of those 2 covariates, we make 4 scatter plots. For each scatter plot, the x-axis is the covariate, and the y-axis is 1 of 4 curve parameters: a, b, c, and also the quantity $a + (1 - a)b$, which is equal to the total initial drop in function value, relative to the pre-treatment function value. This last quantity is labelled as 'drop' in the scatter plots. Trends in these scatter



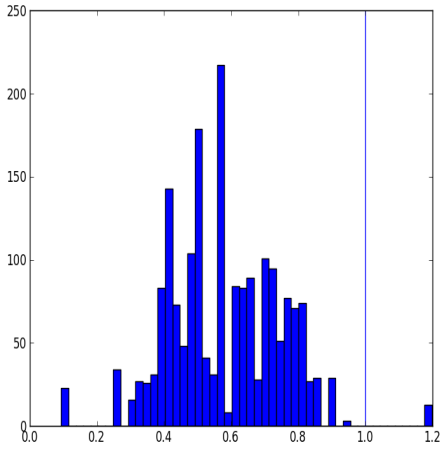
(a) $\phi^a = 0.9$



(b) $\phi^a = 0.5$



(c) $\phi^a = 0.2$



(d) $\phi^a = 0.01$

Figure 4-3: histogram of posterior of B_a when ϕ^a is fixed to various values during inference

plots would lead us to believe that the coefficients in the generalized linear models for the a, b, c parameters would be non-zero.

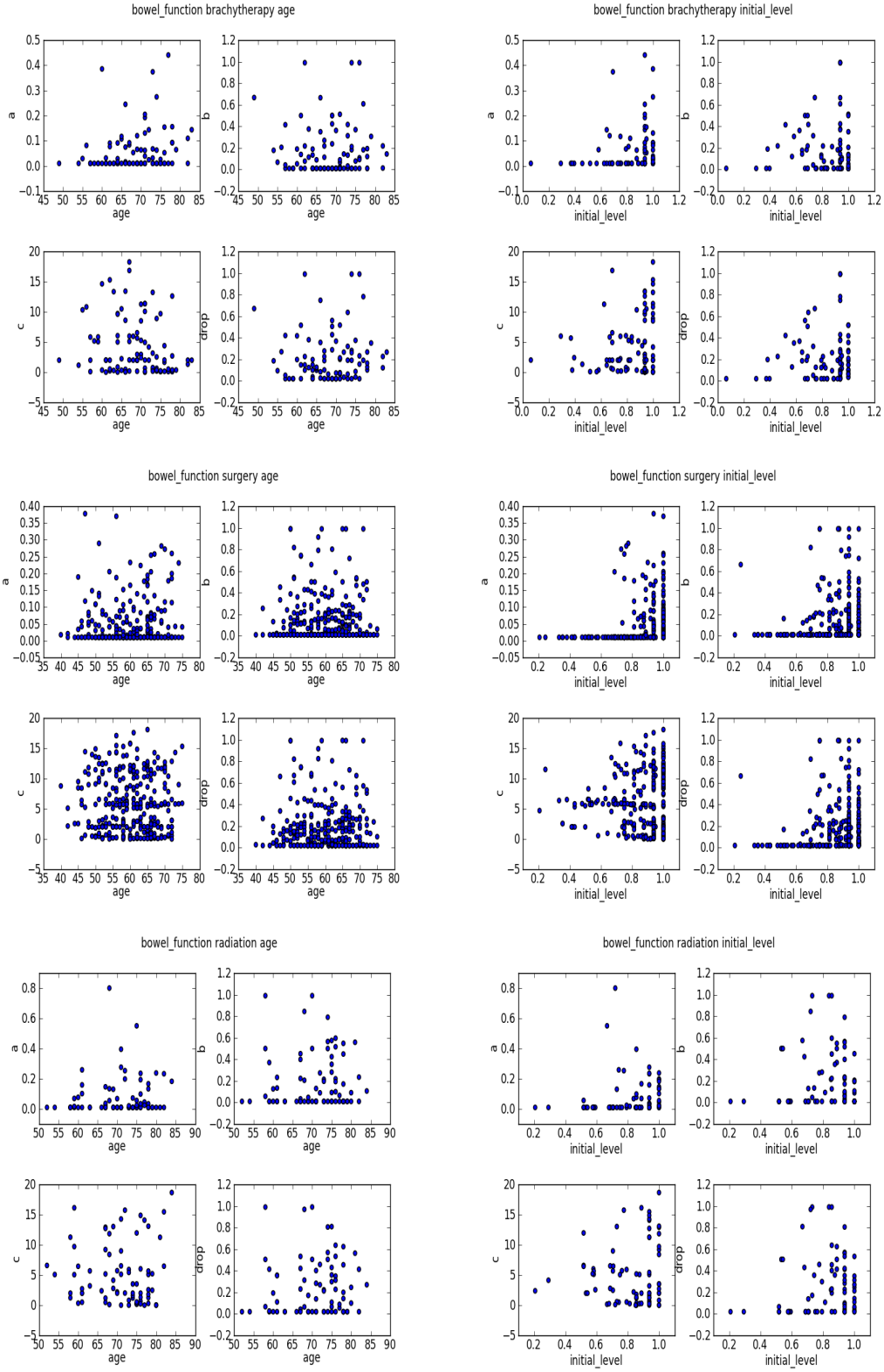


Figure 4-4: Plots of bowel function curve parameters vs initial function level and age attribute, stratified by treatment

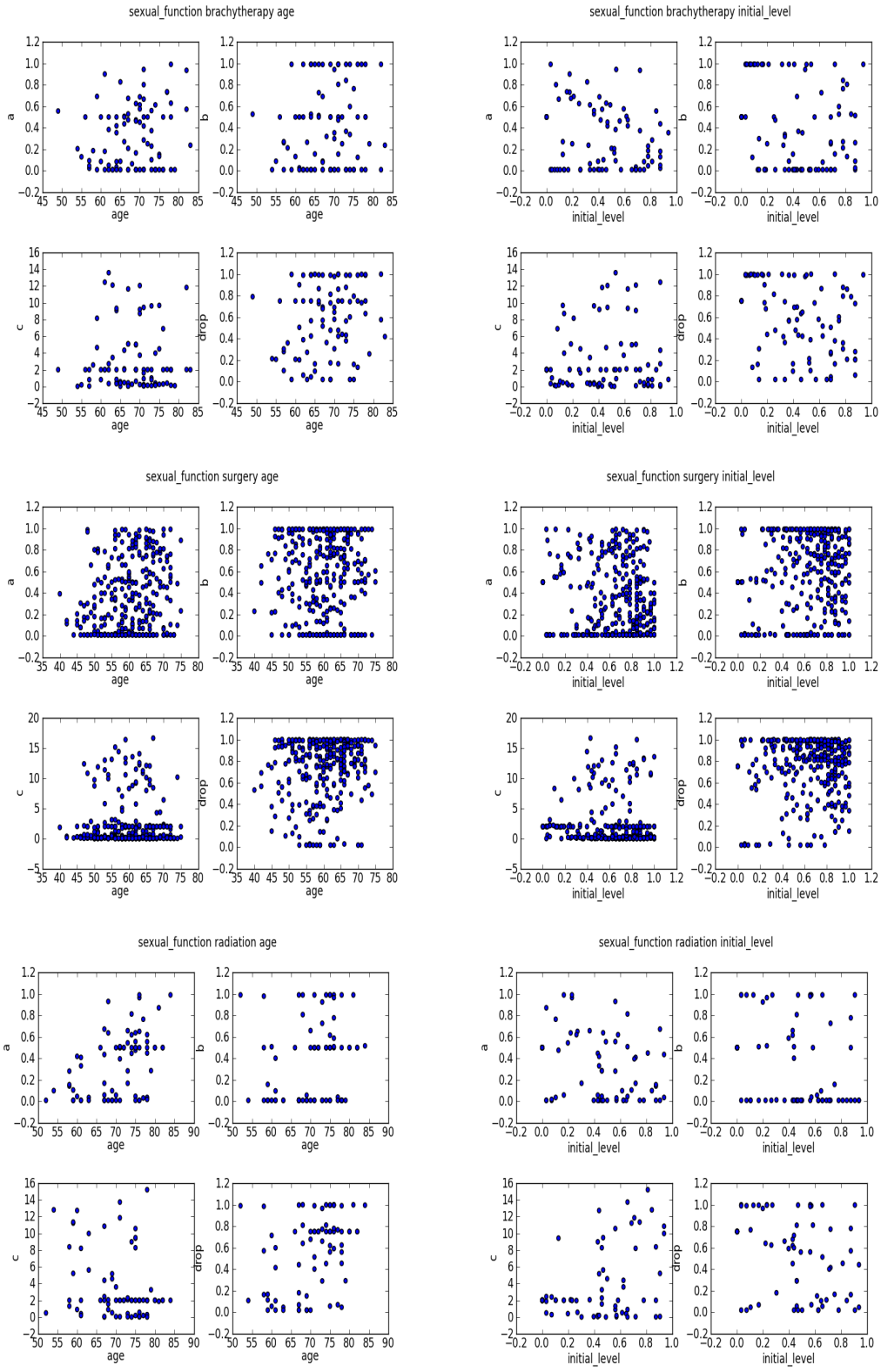


Figure 4-5: Plots of sexual function curve parameters vs initial function level and age attribute, stratified by treatment

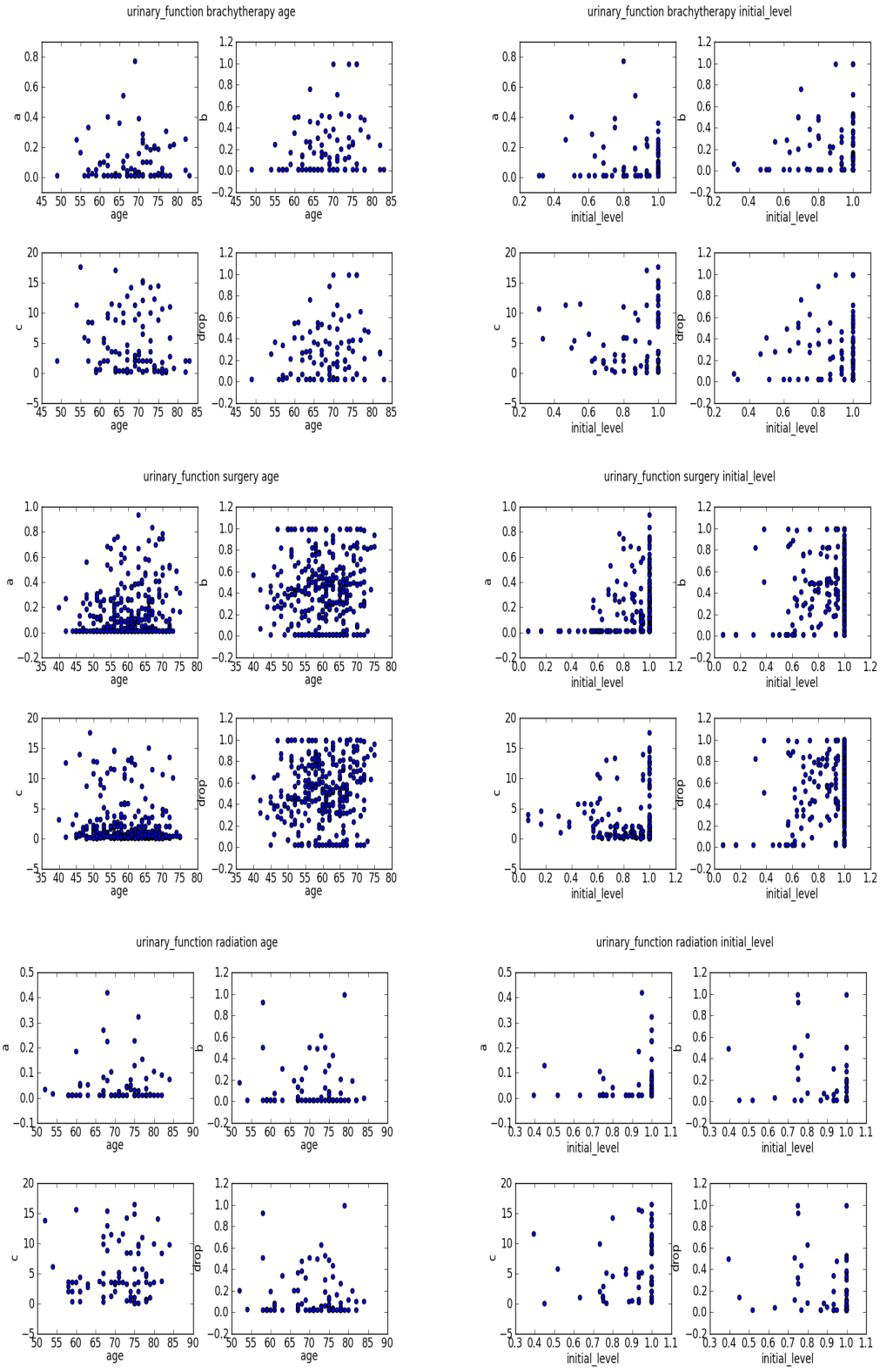


Figure 4-6: Plots of sexual function curve parameters vs initial function level and age attribute, stratified by treatment

Bibliography

[1] Smithson M, Verkuilen J. A Better Lemon Squeezer, Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. Psychological Methods. 2006.

[2] Gore J, Kwan L, Lee S, Reiter R, Litwin M. Survivorship Beyond Convalescence: 48-Month Quality-Of-Life Outcomes after Treatment for Localized Prostate Cancer. J Natl Cancer Inst. 2009.

[3] Gelman A. Bayesian Data Analysis. 2003.

[4] Metropolis N, Rosenbluth A, Rosenbluth M. Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics. 1953.